Science and Technology Indonesia
e-ISSN:2580-4391 *p*-ISSN:2580-4405
Vol. 6, No. 2, April 2021

Research Paper

Diagnosis of Diabetes Mellitus in Women of Reproductive Age using The Prediction Methods of Naive Bayes, Discriminant Analysis, and Logistic Regression

Yulia Resti[1]*, Endang Sri Kresnawati[1], Novi Rustiana Dewi[1], Des Alwine Zayanti[1], Ning Eliyati[1]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Sumatera Selatan, Indonesia*
*__Corresponding author__: yulia_resti@mipa.unsri.ac.id

Abstract
Diabetes is a chronic disease that can cause serious illness. Women are four times more likely to develop heart problems caused by diabetes. Women are also more prone to experience complications due to diabetes, such as kidney problems, depression, and decreased vision quality. Nearly 200 million women worldwide are affected by diabetes, with two out of five affected by the disease being women of reproductive age. This paper aims to predict women with at least 21 years of age having diabetes based on eight diagnostic measurements using the statistical learning methods; Multinomial Naive Bayes, Fisher Discriminant Analysis, and Logistic Regression. Model validation is built based on dividing the data into training data and test data based on 5-fold cross-validation. The model validation performance shows that the Multinomial Naïve Bayes is the best method in predicting diabetes diagnosis. This paper's contribution is that all performance measures of the Multinomial Naïve Bayes method have a value greater than 93 %. These results are beneficial in predicting diabetes status with the same explanatory variables.

Keywords
Diabetes Prediction, Discriminant Analysis, Logistic Regression, Naïve Bayes

Received: 4 March 2021, Accepted: 14 April 2021
https://doi.org/10.26554/sti.2021.6.2.96-104

1. INTRODUCTION

According to World Health Organization (Organization, 2017), diabetes is a chronic disease that can cause serious illness. Women are four times more likely than men to develop heart disease caused by diabetes. Additionally, women are more prone to developing diabetes-related complications such as kidney disease, depression, and vision loss. According to estimates, nearly 200 million women worldwide are affected by diabetes, with two out of five affected by the disease being women of reproductive age. The estimated total number of people with diabetes by the year 2040 will reach 313 million (Nishtar, 2017). Related to the financial burden, the enormous cost of treating diabetes at $ 13,700 per year will double by 2030 (Abdollahi et al., 2019). Preventive action can be done earlier if diabetes prediction can be made correctly and the costs incurred in treating diabetes can be significantly reduced (Federation, 2016).

Predictions of diabetes status (diagnosis) of women of reproductive age have been carried out by (Maniruzzaman et al., 2017; Zou et al., 2018; Tigga and Garg, 2020). Especially (Maniruzzaman et al., 2017) employed many approaches to ascertain a patient's diabetes status. Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, and Gaussian Process (GP) are all included. The accuracy, sensitivity, specificity, precision, and negative predictive value (NPV) of the model were evaluated.

The findings showed that the GP method has the best performance with an accuracy level of 79.94 %, a sensitivity of 91.79 %, specificity of 63.33 %, precision of 84.91 %, and a negative predictive value of 62.50 %. (Tigga and Garg, 2020) also predicted diabetes status (diagnosis) of women of reproductive age. They found the Random Forest method is the best performance comparing Logistic Regression, K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, and Decision Tree, but the accuracy as the performance measured is low of 59.33 %. (Zou et al., 2018) also guarantee the early diagnosis of diabetes. They used a decision tree, random forest, and neural network to find factors that predict hospital admissions, based on a dataset from Luzhou, China. They show that the random forest method produces the highest accuracy while producing an accuracy of 80.84 %. The better the accuracy of a prediction method, the smaller the error rate. All studies still have a significant error rate of around a minimal 20 %. The studies claim that a classification system's

accuracy rate should be 85 % or better (Aronoff, 1985; Liu and Feng, 2020).

Naïve Bayes, Discriminant Analysis, and Logistic Regression methods are prediction methods based on statistical learning, which often have high accuracy. The Naïve Bayes method is also a robust classification algorithm and works well when it has a high dimensional space (Adetunji et al., 2018; Agarwal et al., 2015). The Naïve Bayes method can both discrete and continuous data types. When all the input variables are of continuous type, these variables are assumed to have a multivariate Gaussian distribution, known as Gaussian Naïve Bayes. However, when these variables do not have a multivariate Gaussian distribution, they are discretized so that they are nominal (Soria et al., 2011; Resti et al., 2020). The solution of this case can be obtained using Multinomial Naïve Bayes method (Xu et al., 2017; Lohar et al., 2017). Fisher Discriminant Analysis method intended for input (explanatory) variables that all represent continuous-type data. This method assume the homogeneity of group covariance matrices but does not require multivariate normality (Sever et al., 2005). In contrast to Discriminant Analysis methods, the Logistic Regression method can be implemented on all types of variable input, both discrete and continuous data types. This method is more flexible because it does not have assumptions regarding the distribution of the variables (Hosmer Jr et al., 2013).

Most data sets in the real world often include continuous variables, but generally, these data do not have a Gaussian multivariate distribution. Laboratory measurements related to diabetes diagnosis are generally continuous variables. This study predicts the diabetes diagnosis based on risk factors that can cause women of reproductive age to have diabetes using Naïve Bayes, Discriminant Analysis, and Logistic Regression methods.

2. EXPERIMENTAL SECTION

2.1 Data

This study's data is the Pima Indian dataset, where the object of observation is women aged 21 - 81 years who have individual diagnostic measurements. The group diagnosed with diabetes was given the name "Yes" (34.90 %), while the group that was not diagnosed with diabetes was given the name "No" (65.10 %). Eight diabetes diagnostic measurements are explanatory variables in this work. They are the blood glucose level (mg/dl), diastolic blood pressure (mm Hg), tricep skinfold thickness (mm), serum insulins level (μU/ml), Body Mass Index (kg/m^2), diabetes pedigree function (a function which scores likelihood of diabetes based on family history), age (years), and pregnancies (times).

2.2 Method

The stages of this research, as given in Figure 1 follows:

There are 768 objects of observation in this work. The dataset is randomly divided into five-folds in a similar size, where one-fold is testing data, and the rest of the four-fold is learning data (Lantz, 2013). The five-fold is one of the k-fold techniques that are less biased (Rodriguez et al., 2009). The composition of data for each group, "Yes" and "No" is presented in Table 1.

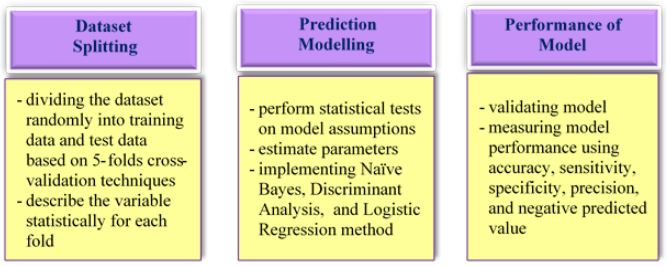

Figure 1. Research Method

The mean and standard deviation for all variables of each fold is given in Table 2.

In the Naïve Bayes method, a patient with k-diagnostic measurements as explanatory variables is predicted to have diabetes if the maximum posterior probability in the "Yes" group is written in (1).

$$P(G_{Yes}|X_1,\cdots,X_k) = \frac{(P(G_{Yes})P(X_1,\cdots,X_k|G_{Yes})}{P(X_1,\cdots,X_k)} \tag{1}$$

Vice versa, a patient is predicted to has not diabetes if the maximum posterior probability in the "No" group as written in (2).

$$P(G_{No}|X_1,\cdots,X_k) = \frac{P(G_{No})P(X_1,\cdots,X_k|G_{No})}{P(X_1,\cdots,X_k)} \tag{2}$$

Equations (1) and (2) become (3) and (4) because the denominator in these equations is a constant.

$$P(G_{Yes}|X_1\cdots,X_k) = P(G_{Yes})P(X_1\cdots,X_k|G_{Yes}) \tag{3}$$

$$P(G_{No}|X_1,\cdots,X_k) = P(G_{No})P(X_1,\cdots,X_k|G_{No}) \tag{4}$$

The likelihood function in equation (3) can be construed using product rule as presented in (5),

$$\begin{aligned} P(X_1,\cdots,X_k|G_{Yes}) &= P(X_1|G_{Yes})P(X_2,\cdots,X_k|G_{Yes},X_1) \\ &= P(X_1|G_{Yes})P(X_2|G_{Yes},X_1)P(X_3,\cdots,X_k|G_{Yes},X_1,X_2) \\ &= P(X_1|G_{Yes})P(X_2|G_{Yes},X_1)P(X_3|G_{Yes},X_1,X_2) \\ &\quad P(X_4|G_{Yes},X_1,X_2,X_3)P(X_5,\cdots,X_k|G_{Yes},X_1,X_2,X_3,X_4) \\ P(X_1,\cdots,X_k|G_{Yes}) &= P(X_1|G_{Yes})P(X_2|G_{Yes},X_1)P(X_3|G_{Yes},X_1,X_2) \\ &\quad P(X_{k-1}|G_{Yes},X_1,\cdots,X_{k-2}P(X_k|G_{Yes},X_1,\cdots,X_{k-1}) \end{aligned} \tag{5}$$

However, definitions such as (5) involve complex and ineffective calculations. The strong-independence assumption called Naïve defines the likelihood function as (6) makes calculations simpler and more efficacious (Alpaydin, 2020),

$$P(X_1,\cdots,X_k|G_{Yes}) = \prod_{d=1}^{k} P(X_d|G_{Yes}) \tag{6}$$

Table 1. Composition of Testing and Learning Data

	Testing Data (one fold)				
Group	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Yes	54	55	51	56	52
No	100	99	103	97	101
Total	154	154	154	153	153
	Learning Data (four folds)				
Group	except fold 1	except fold 2	except fold 3	except fold 4	except fold 5
Yes	214	213	217	212	216
No	400	401	397	403	399
Total	614	614	614	615	615

Table 2. Mean and Standard Deviation of Each Fold

Variable	Statistic	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Blood Glucose Level (X1)	μ	121	121.1	116.4	122.8	123.2
	σ	32.8	30.06	30.96	33.27	32.64
Diastolic Blood Pressure Level (X2)	μ	68.66	69.82	66.29	70	70.78
	σ	22.78	18.97	21.64	16.47	15.85
Tricep Skinfold Thickness (X3)	μ	20.75	18.87	20.08	20.68	22.31
	σ	15.92	15.39	15.23	17.81	15.28
Serum Insulins Level (X4)	μ	77.42	79.2	81.61	78.03	82.75
	σ	111.74	113.58	127.84	109.9	113.65
Body Mass Index (X5)	μ	31.82	31.92	31.62	32.42	32.18
	σ	8.83	7.19	8.44	7.35	7.55
Diabetes Pedigree Function (X6)	μ	0.44	0.5	0.47	0.47	0.48
	σ	0.26	0.34	0.36	0.33	0.36
Age (X7)	μ	32.63	31.95	32.19	35.55	33.91
	σ	11.36	12.39	10.9	12.5	11.34
Pregnancies (X8)	μ	3.92	3.32	3.34	4.41	4.24
	σ	3.59	3.11	2.9	3.42	3.66

This Naïve assumption makes (3) become (7), and with the same assumption (4) becomes (8),

$$P(G_{Yes}|X_1,\cdots,X_k) = P(G_{Yes})\prod_{d=1}^{k} P(X_d|G_{Yes}) \tag{7}$$

$$P(G_{No}|X_1,\cdots,X_k) = P(G_{No})\prod_{d=1}^{k} P(X_d|G_{No}) \tag{8}$$

Each of probability $P(X_d|G_{Yes})$ and $P(X_d|G_{No})$ in (7) and (8) are assumed to have a Gaussian distribution, and the parameters are found using maximum likelihood estimation. In this work, the Naïve assumption was investigated using the Pearson correlation (Chatterjee and Simonoff, 2013),

$$\rho = \frac{n\sum_{i,j=1}^{n} x_i x_j - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j}{\sqrt{(n\sum_{i=n}^{n} x_i^2 - (\sum_{i=n}^{n} x_i)^2) - (n\sum_{j=n}^{n} x_j^2 - (\sum_{j=n}^{n} x_j)^2))}} \tag{9}$$

This work proposed the Multinomial Naive Bayes method (Xu et al., 2017) when the input variables do not have a multivariate Gaussian distribution. In this method, each of probability $P(X_d|G_{Yes})$ and $P(X_d|G_{No})$ in (7) and (8) is defined as,

$$P(X_d|G_{Yes}) = \frac{(\sum_{c}^{m} n_c(X_d|G_{Yes}) + 1}{n(X_d|G_{Yes}) + m} \tag{10}$$

$$P(X_d|G_{No}) = \frac{(\sum_{c}^{m} n_c(X_d|G_{No}) + 1}{n(X_d|G_{No}) + m} \tag{11}$$

the prior probability of each group written as,

$$P(G_{Yes}) = \frac{(\sum_{d=1}^{k} n(X_d|G_{Yes}) + 1}{n + g} \tag{12}$$

$$P(G_{No}) = \frac{(\sum_{d=1}^{k} n(X_d|G_{No}) + 1}{n + g} \tag{13}$$

Table 3. Confusion Matrix

		Actual	
		Diabetic	Control
Diagnosis	Diabetic	True Positive (TP)	False Positive (FP)
	Control	False Negative (FN)	True Negative (TP)

For the "Yes" group, $n_c\,(X_d|G_{Yes})$ is the number of event "Yes" in variable X_d with category c, $n(X_d|G_{Yes})$ is the number of the event "Yes" in all variables X, $n(G_{Yes})$ is the number of events that occurred in the "Yes' group, m is the number of categories in the variable X_d, and g is the total of groups.

Related to the multivariate Gaussian assumption in the Naïve Bayes method, this work proposed the Henze-Zirkler (Székely and Rizzo, 2005) test, respectively as in equations (14) – (19),

$$T_\rho(p) = \varphi - 2(1+\gamma^2)^{-p/2}\theta + (1+2\gamma^2)^{-p/2} \tag{14}$$

where p be the number of variables, n be the total sample size, X_i be the $i-th$ sample, X_l be the $l-th$ standardized sample, $\bar{X}$ be the sample mean vector and $_S$ be the sample covariance matrix, respectively.

$$\varphi = \frac{1}{n}\sum_{i,l=n}^{n} exp(\frac{-\gamma^2}{2}\|H-i-H-l\|^2) \tag{15}$$

$$\gamma = \gamma_p(n) = \frac{1}{\sqrt{2}}(\frac{n(2p+1)}{4})^1/(p+4) \tag{16}$$

$$\|H_i - H_l\| = (X_i - X_l)^T S_n^{-1}(X_i - X_l) \tag{17}$$

$$\theta = \frac{1}{n}\sum_{i=n}^{n} exp(\frac{-\gamma^2}{2(1+\gamma^2)}\|H_i\|^2) \tag{18}$$

$$\|H_i\|^2 = (X_i - \bar{X}_n)^T S_n^{-1}(X_i - \bar{X}_n) \tag{19}$$

The null hypothesis of the inference is that a Gaussian distribution is available for each group's probability. The hypothesis is rejected if the p-value is smaller than the significant level.

Next, this work proposes discriminant analysis to predict diabetes status. This method is often successful because of its ability to provide consistently good results in many cases (Le et al., 2020). Fisher Discriminant Analysis (FDA) being the most venerable variant. FDA presupposes the homogeneity of group covariance matrices but does not require multivariate normality (Sever et al., 2005).

A patient with characteristic $X = (X_d)^T, d = 1, 2, \cdots, k$ is classified into the j^{th} group, j = "Yes", "No", if the linear combination, $Y = V^T X$, is maximum, where (Sever et al., 2005; Ghojogh et al., 2019; Mika et al., 1999),

$$V = S_W^{-1}(\mu_1 - \mu_2) \tag{20}$$

$$S_W = \sum_{j=1}^{2} S_j \tag{21}$$

$$S_j = \sum_{x_i \in j^{th}{}_g} (X_i - \mu_j)(X_i - \mu_j)^T \tag{22}$$

$$\mu_j = \frac{1}{n_j}\sum_{x_i \in j^{th}{}_g} X_i \tag{23}$$

Related to the homogeneity of group covariance matrices in the Discriminant Analysis method, this work proposed the Bartlett tests (Snedecor and Cochran, 1989). For the number of variables (p), number of groups (k), the total sample size (n), and sample size in group $j(n_j)$,

$$Bartlett = \frac{(n-k)lnS^2 - \sum_{j=1}^{k}(n_j - 1)lnS_j^2}{1 + (1/(3(k-1)))((\sum_{j=1}^{k} 1/(n_j - 1) - 1/(n-k))} \tag{24}$$

Furthermore, the authors propose a method that includes assumptions about the explanatory variables, namely the logistic regression (LR) method. This method determines the probability of a patient is diagnosed with a medical status when the probability is the largest (Hastie et al., 2009; James et al., 2013),

$$P(Y = yes|X = x_d) = \frac{\sum_{d=0}^{8} exp(\beta_{yes,d}x_{yes,d})}{1 + \sum_{d=0}^{8} exp(\beta_{yes,d}x_{yes,d}) + \sum_{d=0}^{8} exp(\beta_{no,d}x_{no,d})} \tag{25}$$

$$P(Y = no|X = x_d) = \frac{1}{1 + \sum_{d=0}^{8} exp(\beta_{yes,d}x_{yes,d}) + \sum_{d=0}^{8} exp(\beta_{no,d}x_{no,d})} \tag{26}$$

The parameters of β_d can be determined using the combination of both the maximum likelihood estimation and the Newton-Raphson method.

In this prediction method, simultaneous and partial tests to determine the effect of explanatory variables are the likelihood ratio and Wald test (Hosmer Jr et al., 2013) as presented in (27) and (28). The null hypothesis of each simultaneous and partial test is that each explanatory variable together does not affect the outcome, and the explanatory variable does not affect the outcome. The rejected hypothesis if the p-value is smaller than the significant level.

$$G = -2ln(\frac{(\frac{n_{yes}}{n})^{n_{yes}}(\frac{n_{yes}}{n})^{n_{yes}}}{\prod_{i=1}^{n}(P(Y = yes|X = x_d))^{y_i}(P(Y = no|X = x_d))^{1-y_i}}) \quad (27)$$

$$W = \frac{\hat{\beta}_d}{SE(\hat{\beta}_d)} \quad (28)$$

Furthermore, this work implements the Multinomial Naive Bayes, Fisher Discriminant Analysis, and Logistic Regression methods to predict whether the woman has diabetes.

Next, the four methods' performance to predict the diabetes status are evaluated use measures accuracy, sensitivity, specificity, precision, and negative predicted value (Ghatak, 2017; Burger, 2018) based on the confusion matrix in Table 3.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (29)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (30)$$

$$Specificity = \frac{TN}{(FP + TN)} \quad (31)$$

$$PPV = \frac{TP}{(TP + FP)} \quad (32)$$

$$NPV = \frac{TN}{(FN + TN)} \quad (33)$$

3. RESULTS AND DISCUSSION

This paper proposed three methods to predict diabetes status: Naïve Bayes, Discriminant Analysis, and Logistic Regression. The validation model used the five-fold cross-validation, which is one of the k-fold technique is less biased (Rodriguez et al., 2009; Bengio and Grandvalet, 2004).

First, this study investigated the Naive and multivariate Gaussian distribution assumptions in connection with the Naive Bayes method. The naïve assumption is investigated using the correlation between input variables. Figure 2 depicts the correlations between variables in the first train data. Here, only X_1 and X_8 has a correlation greater than 0.5, while the other correlations

Table 4. Multivariate Gaussian Test

Henze-Zickler Test	Group	
	Yes	No
stat	1.4	2.29
p-value	0	0

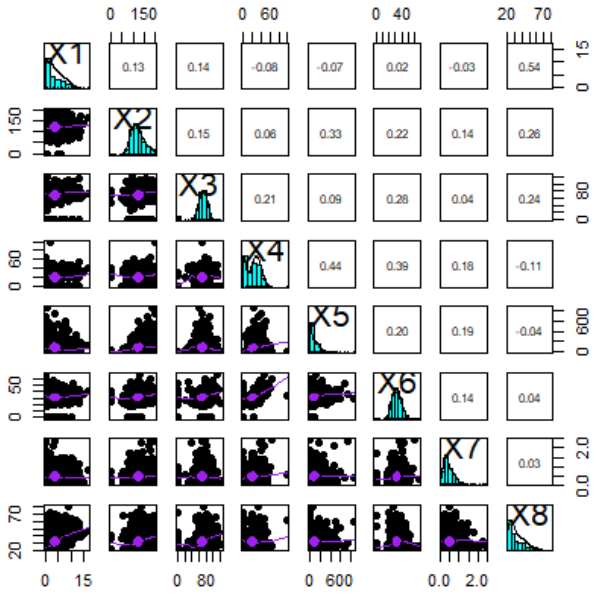

Figure 2. Correlation Between Variables in The First Train Data

Table 5. The Explanatory Variable Discretization

Variable	Discretization	Source
Blood Glucose Level (2 hours after eating)	Normal : < 140 mg/dl High : ≥ 140 mg/dl	(Araki et al., 2020)
Diastolic Blood Pressure	Low: < 60 mm hg Normal: 60-80 mm hg Pra-hypertension: 81-89 mm hg Hypertension: ≥ 90 mm hg	(Tsujimoto and Kajio, 2018)
Tricep Skinfold Thickness	≤ 30 mm > 30 mm	(Marrodán et al., 2015; Khadilkar et al., 2015)
Serum Insulins Level (2 hours after ingestion of sugary drink)	≤ 166 µU/ml > 166 µU/ml	(Zlativa et al., 2017)
Body Mass Index (BMI)	Normal: < 30 kg/m2 Obesity: ≥ 30 kg/m2	(Nuttall, 2015)
Diabetes Pedigree Function (DPF)	< 0.40 0.40 - 0.80 > 0.80	(Survey, 2017)
Age	≤ 35 years > 35 years	(Lampinen et al., 2009)
Pregnancies	≤ 4 times > 4 times	(Karegowda et al., 2012)

are less than 0.5. In four others training data, the correlation also shows the same thing. Only the correlation X_1 and X_8 correlate greater than 0.5; all the rest correlate 0.01 - 0.44 (regardless of positive or negative).

Due to the assumption of a multivariate Gaussian distribution in the Naïve Bayes method, it is not easy to find the actual world data (Hallin and Paindaveine, 2009), especially laboratory measurement data related to the diagnosis of diabetes (Maniruzzaman et al., 2017). Table 4 informs that the input variables in the first train data, either the "Yes" group or the "No" group, have no multivariate Gaussian distribution based on the Henze-Zickler test. The same thing happened in four others training data.

Following that, this work refers to (Xu et al., 2017; Lohar et al., 2017) to address the violation. Table 5 presents the variable discretization for implementing the Multinomial Naive Bayes process.

Next, for the Fisher Discriminant Analysis method, the assumption related to the covariance matrix's homogeneity is given in Table 6 (We exclude the X8 variable because it is not a continuous type). The table shows that the p-value is not smaller than the significance level of 0.05. There is enough evidence to suggest that the covariances of the two groups are homogenous. The results of this test indicate the Fisher Discriminant Analysis method can be implemented in this work.

Furthermore, the Goodness-of-Fit statistics in Table 7 provide two different measures to assess the model's fit for the logistic regression method. However, different results were obtained. The models fit the data well if the p-value is less than 5% based on the deviance measure; the model fits the data best, but based on the Pearson measure, the model fits the data best.

Table 6. The Covariance Matrix Homogeneity Test

Bartlett Test	Group	
	Yes	No
stat	2.36	5.23
p-value	0.12	0.17

Table 7. Goodness of Fit

	Chi-Square	df	p-value
Deviance	720.33	744	0.73
Pearson	811.69	744	0.04

Another choice to get a comprehensive understanding of our model is to look at the statistical parameters presented in Table 8. The final model presents all the variables in the study, plus their overall statistical value. The p-value associated with the full model shows that the full model significantly influences the predicted value in the standardized prediction equation.

Table 9 shows the explanatory variables are statistically significant of the reduced model. The reduced model is formed by omitting an effect from the final model. In this work, the reduced model is equivalent to the final model because omitting the effect does not increase freedom. The likelihood ratio test showed that the skinfold thickness, insulin, pedigree, age, and pregnancies were not the statistically significant variables. From the interim results of the Wald test, the insignificant variables were the same except for pregnancy. People who have diabetes

Table 8. Model Fitting Information

Model	Model Fitting Criteria -2 loglikelihood	Likelihood Ratio Test Chi-Squared	df	p-value
Intercept only	993.484			
Final	720.332	273.151	23	0

Table 9. Likelihood Ratio Test

	Model Fitting Criteria -2 loglikelihood of reduced model	Likelihood Ratio Test Chi-Squared	df	p-value
Intercept only	720.33	0	0	
glucose (X1)	825.46	105.13	1	0
blood pressure (X2)	727.86	7.53	1	0.01
skinfold thickness (X3)	720.47	0.14	1	0.71
insulin (X4)	721.64	1.31	1	0.25
body mass index (X5)	759.85	39.52	1	0
pedigree function (X6)	722.55	2.22	1	0.14
age (X7)	722.88	2.54	1	0.11
pregnancies (X8)	745.97	25.63	16	0.06

Table 10. Prediction Performance using Naïve Bayes

Testing	Accuracy	Sensitivity	Specificity	Precision	NPV
Fold 1	93.51	90.74	95	90.74	95
Fold 2	96.75	92.73	98.99	98.08	96.08
Fold 3	97.4	96.08	98.06	96.08	98.06
Fold 4	95.42	92.86	96.91	94.55	95.92
Fold 5	96.08	100	94.06	89.66	100
Average	95.83	94.48	96.6	93.82	97.01
SD	1.5	3.63	2.06	3.56	2.01

Table 11. Prediction Performance using Fisher Discriminant Analysis

Testing	Accuracy	Sensitivity	Specificity	Precision	NPV
Fold 1	74.03	46.3	89	69.44	75.42
Fold 2	81.82	60	93.94	84.62	80.87
Fold 3	78.57	64.71	85.44	68.75	83.02
Fold 4	77.12	51.79	91.75	78.38	76.72
Fold 5	74.51	59.62	82.18	63.27	79.81
Average	77.21	56.48	88.46	72.89	79.17
SD	3.18	7.34	4.74	8.5	3.09

Table 12. Prediction Performance using Logistic Regression

Testing	Accuracy	Sensitivity	Specificity	Precision	NPV
Fold 1	68.18	32.69	86.27	54.84	71.54
Fold 2	66.88	33.96	84.16	52.94	70.83
Fold 3	69.48	40.74	85	59.46	72.65
Fold 4	67.32	32.73	86.73	58.06	69.67
Fold 5	68.63	35.19	86.87	59.38	71.07
Average	68.1	35.06	85.81	56.94	71.15
SD	1.03	3.34	1.18	2.91	1.08

have more significant health care needs, have fewer productive years, and are significantly more likely to be absent from work. The effectiveness of the process in predicting insulin resistance needs to be confirmed. Methods are used to reduce prediction errors. Preventive action can be done earlier, and the costs incurred in managing diabetes can be significantly reduced by making a screening test available to patients.

In either the statistical or machine learning field, a classification technique's performance is usually measured in a term of prediction error (Rodriguez et al., 2009; Bengio and Grandvalet, 2004). The term can be found by subtracting the total probability by accuracy. However, the term is not enough in medical informatics data. Several other terms are needed to measure classification techniques' performance, such as accuracy, sensitivity, specificity, precision, and negative predictive value (NPV) (Maniruzzaman et al., 2017; Ghatak, 2017; Burger, 2018) as formulated in (24) – (27).

In this paper, we have investigated the performance of two statistical classification techniques based on the five-fold cross-validation technique. The Naïve Bayes and discriminant analysis parameters are estimated using maximum likelihood estimation, while the Logistic Regression parameters use both maximum likelihood estimation and the Newton-Raphson method. The performance measurements of accuracy, sensitivity, specificity, precision, and negative predicted value (NPV) are presented in Table 10 - Table 12.

We have found the diabetes status classification with Multinomial Naive Bayes to be the highest performing. In this case, the Multinomial Naive Bayes method's performance gives accuracy 95.83 %, sensitivity 94.48 %, specificity 96.60 %, precision 93.82 %, and NPV 97.01 % for the Pima Indian diabetes dataset, and this performance is the best compared to other classification methods. Generally, medical data performance, especially disease prediction, should have high accuracy and sensitivity to receive appropriate treatment. The performance of the Multinomial Naive Bayes method in this work also is the best compared to those obtained by (Maniruzzaman et al., 2017; Zou et al., 2018; Sisodia and Sisodia, 2018; Nilashi et al., 2017).

4. CONCLUSIONS

Diabetes is a chronic disease that can cause serious illness. Women are four times more likely to develop heart problems caused by diabetes. Women are also more prone to experience complications due to diabetes, such as kidney problems, depression, and decreased vision quality. Correct diabetes status prediction can help doctors as well as patients for proper treatment planning procedures. Moreover, in particular, women can also try to avoid this disease. Empirical evidence demonstrates that the Multinomial Naive Bayes approach outperforms discriminant analysis and Logistic Regression. Based on different performance factors, including accuracy, sensitivity, specificity, precision, and NPV, we can conclude that naive Bayes classifies diabetes status. Therefore, our recommendation is to use Multinomial Naive Bayes for classifying diabetes data as well as medical informatics data.

REFERENCES

Abdollahi, J., B. N. Moghaddam, and M. E. Parvar (2019). Improving diabetes diagnosis in smart health using genetic-based Ensemble learning algorithm. Approach to IoT Infrastructure. *Future Gen Distrib Systems Journal*, **1**; 23–30

Adetunji, A., J. Oguntoye, O. Fenwa, and N. Akande (2018). Web Document Classification Using Naïve Bayes. *Journal of Advances in Mathematics and Computer Science*, **29**(6); 1–11

Agarwal, S., N. Jain, and S. Dholay (2015). Adaptive testing and performance analysis using naive bayes classifier. *Procedia Computer Science*, **45**; 70–75

Alpaydin, E. (2020). *Introduction to machine learning*. 2nd ed. Massachusetts: Massachusetts Institute of Technology.

Araki, R., T. Yamada, K. Maruo, A. Araki, R. Miyakawa, H. Suzuki, and K. Hashimoto (2020). Gamma-Polyglutamic Acid-Rich Natto Suppresses Postprandial Blood Glucose Response in the Early Phase after Meals: A Randomized Crossover Study. *Nutrients*, **12**(8); 2374

Aronoff, S. (1985). The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering and Remote Sensing*, **51**(1); 99–111

Bengio, Y. and Y. Grandvalet (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, **5**; 1089–1105

Burger, S. V. (2018). *Introduction to machine learning with R: Rigorous mathematical analysis*. O'Reilly Media

Chatterjee, S. and J. S. Simonoff (2013). *Handbook of regression analysis*, volume 5. Wiley Online Library

Federation, I. D. (2016). *Cost-effective solutions for the prevention of type 2 diabetes.* Brussels, Belgium:International Diabetes Federation

Ghatak, A. (2017). *Machine learning with R.* Springer

Ghojogh, B., F. Karray, and M. Crowley (2019). *Fisher and kernel Fisher discriminant analysis: Tutorial.* Manifold Learning and Dimensionality Reduction

Hallin, M. and D. Paindaveine (2009). Optimal tests for homogeneity of covariance, scale, and shape. *Journal of Multivariate Analysis*, **100**(3); 422–444

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media

Hosmer Jr, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied Logistic Regression Analysis.* Applied logistic regression

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning.* Springer

Karegowda, A. G., V. Punya, M. Jayaram, and A. Manjunath (2012). Rule based classification for diabetic patients using cascaded k-means and decision tree C4. 5. *International Journal of Computer Applications*, **45**(12); 45–50

Khadilkar, A., R. Mandlik, S. Chiplonkar, V. Khadilkar, V. Ekbote, and V. Patwardhan (2015). Reference centile curves for triceps skinfold thickness for Indian children aged 5–17 years and cut-offs for predicting risk of childhood hypertension: a multicentric study. *Indian pediatrics*, **52**(8); 675–680

Lampinen, R., K. Vehviläinen-Julkunen, and P. Kankkunen (2009). A review of pregnancy in women over 35 years of age. *The open nursing journal*, **3**; 33

Lantz, B. (2013). *Machine learning with R.* Packt publishing, Birmingham-Mumbay

Le, K. T., C. Chaux, F. J. Richard, and E. Guedj (2020). An adapted linear discriminant analysis with variable selection for the classification in high-dimension, and an application to medical data. *Computational Statistics & Data Analysis*, **152**; 107031

Liu, J. E. and P. A. Feng (2020). *Image classification algorithm based on deep learning-kernel function.* 14. Scientific programming, ID 7607612

Lohar, P., K. Dutta Chowdhury, H. Afli, H. Mohammad, and A. Way (2017). ADAPT at IJCNLP-2017 Task 4: a multinomial naive Bayes classification approach for customer feedback analysis task. *Proceedings of the 8th International Joint Conference on Natural Language Processing*; 161–169

Maniruzzaman, M., N. Kumar, M. M. Abedin, M. S. Islam, H. S. Suri, A. S. El-Baz, and J. S. Suri (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer methods and programs in biomedicine*, **152**; 23–34

Marrodán, M. D., M. G.-M. de Espinosa, Á. Herráez, E. L. Alfaro, I. F. Bejarano, M. M. Carmenate, C. Prado, D. B. Lomaglio, N. López-Ejeda, A. Martínez, et al. (2015). Subscapular and triceps skinfolds reference values of Hispanic American children and adolescents and their comparison with the reference of Centers for Disease Control and Prevention (CDC). *Nutricion hospitalaria*, **32**(6); 2862–2873

Mika, S., G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers (1999). Fisher discriminant analysis with kernels. *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop*; 41–48

Nilashi, M., O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi (2017). Accuracy improvement for diabetes disease classification: a case on a public medical dataset. *Fuzzy Information and Engineering*, **9**(3); 345–357

Nishtar, S. (2017). Shape the Future of Diabetes. *Diabetes Voice*, **62**(1); 23–26

Nuttall, F. Q. (2015). Body mass index: obesity, BMI, and health: a critical review. *Nutrition today*, **50**(3); 117

Organization, W. H. (2017). *Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth.* World Health Organization

Resti, Y., I. Yani, F. Burlian, D. A. Zayanti, and I. M. Sari (2020). Improved the Cans Waste Classification Rate of Naive Bayes using Fuzzy Approach. *Science and Technology Indonesia*, **5**(2); 75–78

Rodriguez, J. D., A. Perez, and J. A. Lozano (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, **32**(3); 569–575

Sever, M., J. Lajovic, and B. Rajer (2005). Robustness of the Fisher's discriminant function to skew-curved normal distribution. *Metodoloski zvezki*, **2**(2); 231

Sisodia, D. and D. S. Sisodia (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, **132**; 1578–1585

Snedecor, G. W. and W. G. Cochran (1989). Statistical Methods, eight edition. *Iowa state University Press, Ames,*

Soria, D., J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, and I. O. Ellis (2011). A 'non-parametric' version of the naive Bayes classifier. *Knowledge-Based Systems*, **24**(6); 775–784

Survey, N. F. H. (2017). *International Institute for Population Sciences.* Mumbai 400 088

Székely, G. J. and M. L. Rizzo (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, **93**(1); 58–80

Tigga, N. P. and S. Garg (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, **167**; 706–716

Tsujimoto, T. and H. Kajio (2018). Low diastolic blood pressure and adverse outcomes in heart failure with preserved ejection fraction. *International journal of cardiology*, **263**; 69–74

Xu, S., Y. Li, and Z. Wang (2017). Bayesian multinomial Naïve Bayes classifier to text classification. *Advanced multimedia and ubiquitous engineering*, **448**; 347–352

Zlativa, S., Atanasova., and Ivanova (2017). Glucose and Insulin Reference Ranges in Oral Glucose Tolerance Test. *International Journal of Scientific Research*, **6**(5); 451–452

Zou, Q., K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, **9**; 515

